Preparing resource discovery for digitized music – an analysis of an Australian application

Jennifer Thomas
QUT School of Information Systems
2 George Street
Brisbane QLD Australia
+(61) 7 31382723
ja.thomas@qut.edu.au

Michael Middleton
QUT School of Information Systems
2 George Street
Brisbane QLD Australia
+(61) 7 31382870
m.middleton@qut.edu.au

Margaret Warren
State Library of Queensland
Cultural Centre, South Bank
South Brisbane QLD Australia
+(61) 7 38429815
m.warren@slq.qld.gov.au

Accepted paper for *Proceedings of the ACM IEEE Joint Conference on Digital Libraries (JCDL'07),* June 18-23, 2007 – Vancouver, BC, Canada

ABSTRACT

This paper examines procedures for the creation and delivery of digital music that are being undertaken by contributors to the National Library of Australia's federated music gateway *MusicAustralia*. The case study discusses access to and preservation of digital material as key drivers of the digitization movement, and compares projects being undertaken worldwide. Also analyzed are the underlying digitization principles and standards, and metadata schemas for the description and exchange of digital objects which facilitate record exchange and improve audience reach.

The paper provides an overview of some individual contributing institutions, however particular focus is placed on the State Library of Queensland's (SLQ) approach to preparing its unique Queensland music collection for digital resource discovery in *MusicAustralia*. A detailed analysis of SLQ's strategy is presented, including its risk management approach to copyright implications, and consideration of infrastructure issues affecting the creation, preservation and online delivery of its digital music objects.

Whilst SLQ's current digital music collection is relatively small, it has become core business of SLQ's Arts and Humanities branch, and the collection will expand with the continued incorporation of music material unique to Queensland into the collection. SLQ has developed a sound foundation for digitization based on widely endorsed principles and standards which should allow this to effectively occur.

Categories and Subject Descriptors

Digital Libraries, collection, dissemination, standards, systems issues, user issues

General Terms

Management, Performance, Design, Economics, Reliability, Standardization.

Keywords

Music; Australia; Digitization; Digital libraries; Audio media; Interoperability; Federated searching; Case study.

INTRODUCTION

The Web has provided many alternatives for online provision of sound. These include a range of online music file sharing facilities such as that initiated by Napster, along with compilations of music put together by publishers such as Naxos. Libraries have had to face the challenge of providing a distribution avenue for their users while coming to terms with ownership, preservation, and collection development issues [12].

Libraries have also had to complement these facilities by addressing the way to make physical documents such as scores and music from their own collections more amenable to network availability. This work considers development in this area in Australia under the umbrella of the *MusicAustralia* [14] gateway with particular reference to issues being addressed by the contributors to the sources referenced through the gateway.

The infrastructure for *MusicAustralia* has been established cooperatively on a national basis by means of a collaborative mechanism that enables institutions to make available musical information through a central discovery service. Its success depends both upon the effectiveness of this service and the ability of contributors to contribute effectively to such a federated service [4, 7].

Worldwide, a number of libraries have made digital surrogates of culturally or socially important music collections available online via dedicated websites or federated services. Digital content includes material as varied as images, photographs, sound recordings, manuscripts, musical scores and objects. With the proliferation of information technologies and the relative ease with which such technologies can be acquired and implemented, the digital movement is intensely discussed and has evolved within libraries for a number of reasons. Arguments for access and preservation have driven the trend towards digitization and consequently libraries of all sizes have implemented digitization projects on varying scales, for the purpose of showcasing their collections to an infinitely larger online audience.

Many libraries are also keen to exhibit their online digital collections for the perceived value of the content to different communities of users. Technical advances further enable a value-laden experience for users, for example the *Maine Music Box*, an initiative of a public and academic library cooperative in Maine, USA, allows users not only to view images of musical scores and

listen to sound recordings, but also to manipulate the arrangement of selected works by altering key signatures and instruments used [11].

Additionally, many institutions are forming partnerships such as the one in Maine, in order to share resources and increase budgets. For example, Colorado's *Collaborative Digitization Project* (CDP), an initiative of the Colorado State Library, endeavors, through cooperative efforts with partners, to make digital content about American culture available online [6]. There is a strong focus on collaboration and best practices regarding digital collection development and access, and a strong consideration of the expected usage in selecting material for digitization. As digitization projects can be expensive and resource-intensive, duplication of effort in digitization and resource description is minimized through cooperative and federated services such as the *Maine Music Box*, CDP and *MusicAustralia*.

Libraries can also gain renown and an increased profile as users access their digital collections remotely [10]. Larger projects such as *Collect Britain*, the British Library's digitization initiative, and the Library of Congress's *American Memory*, enjoy larger budgets and are consequently able to digitize widely, whilst espousing free and open access to their digital collections.

1. BACKGROUND

Participants in *MusicAustralia* describe their resources using either the MARC21 format or the Metadata Object Description Schema (MODS) [9]. MODS records are harvested by the National Library of Australia (NLA) which provides the aggregator service. NLA converts MODS records to MARC21 format and loads them into the National Bibliographic Database (NBDB) which it maintains. From this, they are re-selected for presentation in the *MusicAustralia* service. The choice of MODS has enabled collections not described in MARC21 to be included in a trusted aggregation [5].

Using this process, the number of music information resources in *MusicAustralia* has grown to approximately 167,000, of which many are only physically available in the contributors' collections either as print material or sound recordings. Of this total however, approximately 13,000 resources are available online.

1.1 Participants providing digitized material

MusicAustralia, developed by the NLA and the National Film and Sound Archive, harvests music records from libraries, archives, museums, universities and other specialist online music services such as *Australian Music Online* (AMO). However, *MusicAustralia* can only retrieve records from institutions which have the capacity to be able to exchange records with other services and to make their records available for harvesting in specified formats [15]. As previously mentioned, whilst many of the resources found in *MusicAustralia* are not available online, the pool of digital resources such as sound recordings and digital scores held within the service is slowly increasing, due to dedicated digitization projects being carried out by certain contributing institutions.

The NLA for example, operates a digital music initiative making sound recordings and digital scores of unique Australian music available online. Additionally, a number of state libraries in Australia are carrying out similar projects where musical material unique to their jurisdiction is digitized and subsequently harvested by *MusicAustralia*. Thus, with libraries only digitizing material which is unique to their jurisdiction and generally not digitizing material which is already digitized and available elsewhere, the instances of duplicate digital records in *MusicAustralia* is greatly minimized. The NLA and state libraries of Australia are major contributors to *MusicAustralia*.

The National Film and Sound Archive, co-developer of *MusicAustralia* is also a major contributor, especially of online sounds. The role of the Archive is to preserve and share Australia's moving images and sound, including both older and more modern material. Of the Archive's 509 resources in *MusicAustralia*, approximately 490 are audio files and of these approximately 350 are available online.

Australian Music Online (AMO), another major contributor to *MusicAustralia*, is a website which provides for marketing and promotion of new Australian contemporary music. This site arose from an initiative by the Australia Council arising from 1998 Federal Government funding to support music industry development. The AMO web resource [1] contains an extensive amount of content and data about new Australian music. This content is created by AMO as well as aggregated from partner sites featuring Australian music content.

Typically single tracks from released CDs (the CD title is what appears in the catalogue) and the holdings for the entire CD may well be in a number of physical collections from the NBDB.

Due to copyright and budgeting restrictions, and the relatively new practice of digitizing music collections, the digital projects carried out by the contributing institutions are still rather small in scale. Whilst advances in technology allow digitization to rapidly occur, the barriers to access are ultimately social, legal and economic in nature rather than technical [10]. As the digital collection within *MusicAustralia* grows, the contributors hope to find solutions to these barriers that are fair for the artists, copyright holders and users of the service [15].

1.2 Principals and standards for contributors

General digitization protocols for *MusicAustralia* contributors are based on principles aligned with those benchmarked by the Australian Government Information Management Office (AGIMO) in their better practice checklist for digitization [3]. These principles ensure that selected digital material:

- Supports collection development policy
- Remains persistent and accessible over generations of technology migrations
- Has been digitized to a recognized, high quality standard or best practice
- Is exchangeable and interoperable
- Is described using appropriate descriptive and administrative metadata, facilitating successful management of the material
- Can potentially operate as both preservation and access mediums

The standards in place for metadata and digital capture and delivery, which facilitate record exchange and improve audience reach, appear to be the standards generally accepted and adopted worldwide. There are numerous benchmarking publications similar to AGIMO's better practice checklist guiding the management of digital projects, the outcomes of which globally demonstrate or support the above.

1.3 Metadata schemas

NLA piloted *MusicAustralia* using Dublin Core as the initial metadata schema for record description. However, Dublin Core's broad scope of potential description, and lack of content indicators for record creation caused problems for *MusicAustralia* with both the volume and types of records being harvested. Dublin Core could not effectively accommodate the potentially rich description of a digitized music score or sound recording, nor represent both the horizontal and vertical relationships regularly apparent between works or different versions of the same work [13]. This discovery was made early in the project and consequently fuelled the change to the MARC and MODS schemas currently employed by many of the contributing institutions.

Supported by the Library of Congress, MODS utilizes a language-based tag format rather than a numeric tag format, allowing flexible creation of original descriptive records for a wide range of digital objects [9]. MODS records are also readily translatable into MARC21 format. MODS is endorsed by AGIMO and is the recommended exchange schema for *MusicAustralia* contributors [15].

Mandatory record fields for a MODS record in *MusicAustralia* include:

- An Australian content indicator – a geographic code enclosed in the <subject> tag,

AND

- Resource type – either:
    - Notated music
    - Manuscript music
    - Sound recording – musical

OR

- At least one Library of Congress Subject Heading (LCSH) music related heading

The possible combinations of these enable *MusicAustralia* to identify and extract the Australian music records from contributors. Additionally, MODS supports the location of URLs of digital objects in either the <location> or <identifier> tags, redirecting users of *MusicAustralia* to the original record location at the contributing institution.

1.4 Research method

An investigation is underway that includes case studying a number of the contributing agencies to *MusicAustralia* utilizing a case study protocol, combined with analysis of database contributions and elements and modes of operation. Regular examination of the web sites and digitization policies and practices of major contributors to the *MusicAustralia* service is being undertaken.

This paper proposes methods for preparing digitized music for resource discovery and provides a snapshot of current practices by reporting on one of these cases, undertaking an investigation of input procedures, and infrastructure supported at the State Library of Queensland (SLQ). The case study protocol has also been supported by interviewing key personnel at SLQ, and examining benchmark policy and planning documentation developed by statutory bodies, which inform SLQ's approach to music digitization.

2. DIGITIZED SOUND SOURCES

2.1 Intellectual property clearances

SLQ digitizes Queensland music material published prior to 1930 which is free from copyright restrictions. However, SLQ strives to ensure that material with important social value is still publicly accessible by seeking clearance to digitize newer material still within copyright. Nevertheless, copyright clearance for musical works is problematic, and in many cases copyright needs to be cleared for, but is not limited to, lyrics, music, images, published versions and arrangements. On average, it takes SLQ five hours to clear copyright for one piece of music.

The management of "orphan works", i.e. works for which there is no identifiable copyright owner, constitutes an issue for SLQ that continues to remain under consideration. It adopts a risk management approach to the digitization of such works based on the commercial and historical value of the material. Works with low commercial, yet high social value may be selected for digitization. Hudson and Kenyon [8] provide useful commentary on risk management and orphan works which has also informed SLQ's approach. Australian legislation is yet to address the issue specifically, and collecting societies in Australia such as the Australasian Performing Right Association (APRA) and the Copyright Agency Limited (CAL) are yet to formulate mechanisms for such material.

SLQ has also partnered with the Queensland Conservatorium of Griffith University to boost the number of digital sound recordings in its collection. The Conservatorium records staff and student performances of unique Queensland music and retains the copyright on these recordings. SLQ is then offered a non-exclusive, royalty-free license to stream these recordings via its website, and for use at SLQ functions and events.

2.2 Infrastructure for creation and delivery

SLQ plans to continue developing its digitized music collection with the inclusion of more recently published material [17]. As its digital collection grows, estimations of growth and storage requirements depend on SLQ's acquisition strategy, digitization protocol and technical specifications. As a legal deposit repository for unique Queensland material, SLQ has the responsibility to ensure that its collections, both physical and digital, are appropriately cared for and stored to an archival standard [16]. For the digital collections, this means utilizing lossless, uncompressed formats for master files that prevent loss

of data quality during file exchange, technological change and storage migrations.

SLQ creates its preservation digital sound recordings as archival quality broadcast WAVE (.wav) master files. The broadcast WAVE file format, recommended by the International Association of Sound Archives is a preferred format widely adopted by both the recording industry and sound archives and is also the file format recommended by AGIMO. The WAVE format ensures that digital reproductions are as faithful to the original as possible, and is therefore useful as a preservation format. However, to ensure the stability and digital survival of the files through cycles of technological and format obsolescence, these high quality preservation files require high levels of bandwidth and storage.

For digital reproductions from paper such as musical scores and manuscripts and photographic media such as negatives, SLQ utilizes the TIFF format (.tif) as the uncompressed master archival file format. TIFF is widely used in library digitization projects both in Australia and overseas, and is endorsed in the AGIMO better practice checklist as a recommended preservation format. Similar to the WAVE format, the TIFF format also requires high levels of storage but produces faithful image reproductions which are suitable as preservation master files.

Sound files require exponentially more storage than image files, with one minute of CD quality sound necessitating approximately 10MB of storage [15]. However, due to the relatively small size of the current digital music collection at SLQ, and with the narrow scope of material suitable for digitization indicating a slow, measured rate of expected collection growth, storage requirements are not onerous for SLQ.

Endeavor Information Systems' *ENCompass* product is the content management system utilized by SLQ in the management of their digital objects. Records are created for the individual objects and contain descriptive, as well as technical and administrative metadata which facilitate object retrieval, and each object is assigned a unique identifier corresponding to SLQ's file naming conventions [18]. Digital sound records also contain URL links to the different audio formats, e.g. .wav, .mp3, .wma, .ram, relative to the user interface. SLQ has also developed navigation software in-house, which allows users to navigate through digital images attached to certain musical scores.

SLQ's music digitization program, which was initially funded as an Online Public Access in Libraries (OPAL) project, has since become core business of SLQ's Arts and Humanities branch. Cataloguers are employed for the creation of the metadata records, whilst staff from the Microfilm Unit create digitized scores from SLQ's original sheet music holdings. Arts and Humanities staff are also involved in decision-making for collection development [18].

2.3 Online delivery

SLQ has a responsibility to collect, preserve and provide all people in Queensland, including those in regional, rural and remote areas with access to unique Queensland material. This is a key driver in the selection of online delivery formats for its digital collections.

As previously mentioned, SLQ's selected formats for digital masters include the broadcast WAVE format for sound recordings and the uncompressed TIFF format for images. These formats ensure that the digital reproductions are as faithful to the original as possible, and are therefore necessary as preservation formats. However, the amount of storage and bandwidth required makes them unsuitable for delivery over the Web.

SLQ's selected formats for digital audio delivery include dialup- and broadband-compatible access files deliverable via either *Windows Media Player* (.wma) or *RealPlayer* (.ram). This echoes NLA's recommendation of offering more than one online delivery option and thus increasing the end user's control over their online experience. Additionally, these current standard formats for Web delivery of sound ensure that the storage requirements for file access copies are minimal. For example, the *Real Media* codec used at SLQ to compress sound files during the digitization process can produce files reduced to 120 times the size of the original.

For online access copies of image files, the JPEG format (.jpg) is utilized for its efficient compression properties [18]. JPEG files are produced from the TIFF masters. SLQ also uses PDF format to provide downloadable and printable versions of documents.

SLQ's adoption of these online delivery formats not only eases storage requirements, but also maximizes the potential user base of the service by accommodating those with slower modems and smaller download quotas and supporting users with dialup Internet connections, particularly in regional, rural and remote communities.

3. DISCUSSION

The relatively small size of SLQ's current digital music collection and its predicted slow rate of growth due to copyright limitations prevent a backlog of material prioritized for digitization, freeing up time to be spent by staff on quality control, the development of standards, and reconciliation of digitization policy and practice with predicted changes in technologies.

At SLQ, a risk management approach to music digitization is deemed most appropriate, however with the problematic intricacies of copyright law, alternative methods for addressing copyright within libraries and cultural institutions are being developed and trialed internationally. For example, the United States is developing legislation that specifically addresses the management of orphan works [2]. In Australia, the federal Attorney-General's Department may review copyright legislation in 2007 with respect to orphan works. Copyright implications further fuelled SLQ's decision to provide streamed rather than downloadable versions of audio material over the Web. Whilst SLQ's approach to copyright can inform approaches taken by other institutions for similar projects, individual institutions need to perform their own risk assessments and implement copyright management frameworks which sufficiently address their own requirements.

The selection of suitable metadata schemas for the description and exchange of digital objects is also an important process for institutions considering digitizing audio material, and SLQ's retrospective conversion from Dublin Core to MODS is a useful example to consider. Dublin Core, the initial metadata schema for records in *MusicAustralia* posed problems for record exchange due to its broad possibilities for description. As Dublin Core does not provide content indicators narrow enough in scope to harvest

records with the specific information required, *MusicAustralia* converted to MARC and MODS, of which SLQ adopted the latter. Whilst SLQ's task of converting records from Dublin Core to MODS was easy due to the small number of affected records, metadata is an important area to get right and such an example should be considered by institutions in their selection of metadata schemas for implementation.

4. CONCLUSION

With the social shift and increased user base brought about by the growth and capability of the Web, SLQ must continually address the challenges posed by its digital initiatives to ensure that its key roles and responsibilities are supported. As SLQ core business, its digital music collection will expand with the continued incorporation of music material unique to Queensland into the collection, and SLQ has developed a sound foundation for digitization based on widely endorsed principles and standards which should allow this to effectively occur.

SLQ's risk management approach to copyright is also giving socially and culturally important Queensland music a platform for access. The adopted MODS metadata schema facilitates the effective exchange of records between SLQ and the NLA, further exposing material in SLQ's collection to a larger number of users. The selected formats for Web delivery achieve this also, accommodating users with slower modems and smaller download quotas. The selected formats for both preservation and access of digital material have also addressed the infrastructure requirements such as storage and digital survival, and will ensure compatibility with future technological formats or migrations. Ultimately, the digital music material held within SLQ's collection is compatible with *MusicAustralia* in terms of metadata, digital capture and delivery, ensuring serendipitous resource discovery for an infinitely wider audience base.

5. REFERENCES

1. *AMO: Australian Music Online*, nd. Retrieved: 8 January 2007 from: http://www.amo.org.au/default.asp
2. American Library Association. *Copyright: Orphan Works*, 2006. Retrieved: 14 October 2006 from: http://www.ala.org/ala/washoff/WOissues/copyrightb/orphanworks/orphanworks.htm
3. Australian Government Information Management Office. *Digitisation of records: better practice checklist*, 2004. Retrieved: 30 November 2006 from: http://www.agimo.gov.au/practice/delivery/checklists/digitisation
4. Ayres, M.L. *MusicAustralia: building on national Infrastructure*, 2004. Retrieved: 12 April 2007 from: http://www.nla.gov.au/nla/staffpaper/2004/ayres1.html
5. Campbell, D., Easy to do: A brief history of federated harvesting in Australia. in *ETD 2005: evolution through discovery; 8th International Symposium on Electronic Theses and Dissertations*, (Sydney, Australia, 2005) Retrieved 8 January 2007 from: http://adt.caul.edu.au/etd2005/papers/039Campbell.pdf
6. Collaborative Digitization Program. *Welcome to the Collaborative Digitization Program*, 2006. Retrieved: 30 October 2006 from: http://www.cdpheritage.org/cdp/index.cfm
7. Holmes, R. and Ayres, M.L. *MusicAustralia: towards a national music information infrastructure*, 2004. Retrieved: 12 April 2007 from: http://www.nla.gov.au/nla/staffpaper/2004/ayres2.html
8. Hudson, E. and Kenyon, A.T. *Copyright and Cultural Institutions: Guidelines for Digitisation*, 2006. Retrieved: October 9 2006 from: http://ssrn.com/abstract=881699
9. Library of Congress. *Metadata Object Description Schema (MODS)*, 2006. Retrieved: 8 January 2007 from: http://www.loc.gov/standards/mods
10. Lopatin, L. Library digitization projects, issues and guidelines: A survey of the literature. *Library Hi Tech*, *24* (2). 273-289.
11. Lutz, M. The Maine music box: a pilot project to create a digital music library. *Library Hi Tech*, *22* (3). 283 - 294.
12. Maple, A. Online music services and academic libraries. *ARL Bimonthly Report* (244). 11-16.
13. National Library of Australia. *About MusicAustralia*, 2006. Retrieved: 4 October 2006 from: http://www.musicaustralia.org/apps/MA?function=authoredContent&name=about&forceNewTrail=true
14. National Library of Australia. *MusicAustralia*, 2002. Retrieved: 12 April 2007 from: http://www.musicaustralia.org/
15. National Library of Australia. *MusicAustralia: Guidelines for contributors*, 2006. Retrieved: 4 October 2006 from: http://www.musicaustralia.org/apps/MA?function=authoredContent&name=contributors
16. State Library of Queensland. *Legal Deposit*, 2006. Retrieved: 30 October 2006 from: http://www.slq.qld.gov.au/serv/pw/legdep
17. State Library of Queensland. *Music Queensland - About us*, 2006. Retrieved: 2 October 2006 from: http://www.musicqld.slq.qld.gov.au/home/about
18. Young, S. *State Library of Queensland - Digital Standard 2 - Digital capture, format & file naming, version 2.03*, 2006. Retrieved: 16 October 2006 from: http://www.slq.qld.gov.au/about/pub/pol